DEVELOPMENT OF A PATTERN ANALYSIS TECHNIQUE FOR USE IN THE SELECTION OF PREDICTORS

Burton Fredrick Folce

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

DEVELOPMENT OF A PATTERN ANALYSIS TECHNIQUE FOR USE IN THE SELECTION OF PREDICTORS

by

Burton Fredrick Folce, Jr.

Thesis Advisor: R. A. Weitzman

June 1973

DEVELOPMENT OF A PATTERN ANALYSIS TECHNIQUE FOR USE IN THE SELECTION OF PREDICTORS

by

Burton Fredrick Folce, Jr.
Lieutenant Commander, United States Coast Guard
B.S., United States Coast Guard Academy, 1964

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

ABSTRACT

This study describes a computerized item-selection program called PAIN that uses a pattern-analysis approach to select a most-valid subset of items from a set. The results of this study indicate that PAIN is capable of selecting a small subset of items which, when scored by pattern analysis, has greater validity than the original set. It appears that, as well as reducing the sizes of standard tests without losing predictive value, PAIN may also be of value in selecting biographical items of information for use as predictors.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENT

The author wishes to thank professor R. A. WEITZMAN for the role he played in the development of this study. Had it not been for his belief in the value of a selection procedure such as that used by PAIN, and his role as a catalyst, the desire to undertake this study would not have existed. Had it not been for his guidance and knowledge in the area of psychometric measures, this study could not have been completed.

I. INTRODUCTION

Testing has played, and in all likelihood will continue to play, a major role in the classification and selection processes of both industry and the military. The vast amounts of time and money expended in this area warrant investigation of any methods which might increase the efficiency of the techniques involved or improve the end results.

It was the objective of this study to investigate one such method.

II. NATURE OF THE PROBLEM

Testing has played an important role in the military system of classification and placement for many years. Basic schooling assignments and eligibility for promotion are just two of the more important areas that have been greatly influenced by testing and test interpretation. Yet, on many occasions the critical time element involved in testing and the lack of quality information available from tests have combined to make classification and placement a haphazard affair.

The U.S. Navy has taken steps to reduce the magnitude of the testing problem by the development of a computer program called SEQUIN (SEQUential Item Nominator). As a result of the use of SEQUIN it has been shown not only that

the size of a test may be reduced without loss of validity, but also that validity may actually be increased by using a specially selected subset of questions from the original test. In some cases as few as seven items from a test were found to provide information equal to or better than that provided by the complete test. This being the case, it appeared that pattern analysis of a few selected items from a test might be feasible.

The problem of using pattern analysis on a test even as small as 30 items is one of shear size. A test of 30-item size yields over a billion possible patterns. The evaluation of this number of patterns is a formidable job for even a computer, not to mention the problem involved with interpretation of individual results once all the patterns have been evaluated. In fact, in order to establish a predictor value for each pattern that could be encountered, at least a billion subjects would have had to already have taken the test under consideration.

Reducing the size of a test to seven items means that only 128 patterns have to be analyzed. The number of patterns involved is found by raising the number 2 to the power indicated by the number of items in the test. A subset of seven-item size would thus be suitable for pattern analysis.

The objective of this study was to devise and evaluate a method of selecting items from a test that would optimize the validity of the subset selected when scored by pattern analysis.

III. DEVELOPMENT OF A SOLUTION

A. SUBJECTS OF THE STUDY

The records of approximately 2,400 U.S. Navy enlisted men who had attended the Electronics Technician School at San Diego, California, after taking the Electronics Technician Selection Test (ETST) were used as the source data of this study. The validation sample consisted of the first 1,500 subjects in the records who had completed the course of instruction and been assigned a final school grade. The cross-validation sample was composed of the next 750 subjects who met the completion and final-grade assignment requirements.

The ETST is made up of three parts totaling 70 items. Part I consists of 20 items designed to test the subject in the area of mathematics. Part II is of 20-item length also and is related to science. Part III consists of items directed at testing knowledge in the area of electricity and radio and has 30 items in it.

Each of the items on the ETST was treated as a predictor variable to be compared with the criterion of final school grade at the Electronics Technician School.

The computer programs used in this study were written in the FORTRAN language and run on the IBM 360 computer at the U. S. Naval Postgraduate School, Monterey, California. The INTEGER*2 numbering convention was used where possible in

programing to conserve core storage area. The increased time involved in running the program with the use of this convention was not considered critical for this study.

B. DATA CONVERSION

The program to select items for pattern analysis was developed on the premise that all items of the set being considered could be expressed in the form of a "yes-no" or "correct-incorrect" answer. This simplified the programing by allowing the item responses to be handled in a binary form.

The conversion of the raw data was not suitable to a manual method of handling because over 168,000 responses required coding. The conversion was done by using the conversion program shown in the COMPUTER PROGRAMS section(p. 31). This program facilitated the handling of the large volume of information. Most of this program is unique to the situation imposed on the author by the form of the data available. However, the comments contained within this program provide a guideline to the steps required in converting data regardless of the nature of the data.

C. PAIN

The author desired to develop a computer program, which was to be called PAIN (Pattern Analysis Item Numinator), that would select a subset of items from the ETST. SEQUIN could already select a subset of items from the ETST but in a way different from that proposed by PAIN. PAIN was based

on the belief that the pattern of responses could contribute more to the overall value of a predictor than was presently being obtained through the use of SEQUIN or any other method. To do this it was necessary for PAIN to be able to assign scores to each of the possible response patterns associated with a subset of items. The score assigned to a pattern in pattern analysis is the mean score of all subjects in a sample who have that pattern. Once a correlation coefficient was determined for a given subset, it would then be necessary to compare this coefficient with that obtained through the examination of every other subset of the same size available from the main set. This was impractical for reasons which will be explained and an alternate approach was necessary if PAIN was to be used.

The number of different subsets of N items that can be formed by a 70-item set is expressed as the combination of 70 items taken N at a time. This meant that to investigate a subset as small as three items in size would have involved the examination of 54,740 possible subsets, each of which contained eight patterns of response. From the information available concerning SEQUIN it appeared that a subset of seven items would be necessary, at the least, if improvement was desired over the methods presently available.

A seven-item subset would allow the 1,500 subjects of the validation sample to be placed in the 128 response patterns involved with an average distribution of slightly

less than 12 subjects per pattern. This number was felt to be sufficient to establish a fairly stable mean score for each pattern. A second advantage of using seven items in the subset was that it would allow ready comparison with the work of Lieutenant K. Weinberg(personal communication). Lieutenant Weinberg had used the same raw data to investigate the validity of the seven items from the ETST selected as the best predictors by SEQUIN. Unless a reasonable alternative to the examination of all possible response patterns was taken, however, this would have meant the investigation of over 77 trillion patterns, a job that was beyond even a computer approach. This was just for the selection of the seventh item of the subset!

In order to overcome the problem of size, the assumption was made that once an item had been selected as the best for a subset of given size, it would continue to be a part of any larger subset. This allowed the author to say that the item selected as the best item for the subset of one item would be a part of the subset of two items, both of which would be part of the subset of three items, etc. This same approach is used in both stepwise regression and SEQUIN, and would reduce the selection process for the seventh item to an examination of slightly over 8,000 patterns. After PAIN was operative, tests were made to determine the effects of the item-retention assumption on the overall validity of the solution.

To test the item-retention assumption, subsets of two items each were selected randomly from each of the three parts of the ETST to act as the two-item subset in the PAIN program. The program was then allowed to select the items for the completion of the subsets of six-item size. The validities of these subsets were then compared with the validity associated with the selection, by PAIN, of all the items in a subset. The two forced items were selected from individual parts of the ETST rather than the total ETST because the results of the unrestricted selection by PAIN indicated that certain sections of the ETST were more valid than other sections.

PAIN operated by computing mean criterion scores for each pattern of responses in a given subset, assigning these scores to subjects having that pattern of responses, and correlating assigned scores with the subjects' final school grades. PAIN provided the following information when run:

1. Validities of all subsets examined.
2. A list of the items that form the most valid subset of a given size.
3. The validity of the most valid subset of each size.

The final form of PAIN is contained in the COMPUTER PROGRAMS section(p. 31). Representative run times and core storage areas for this program on the IBM 360 computer are contained in Table 2 (p. 22). Details on the roles of important variables and how this program can be adapted for general use are contained in APPENDIX A.

D. CROSS-VALIDATION

That program which the author calls "cross-validation" is in fact a combination of two separate programs. The first section of the cross-validation program was written to obtain mean scores for patterns of responses to items selected from the validation sample. This was done in the validation program but could not be output because it was not known while the program was running which subset would eventually be wanted. Since the score for each pattern of response changed whenever a new item was examined, it would have been necessary to store all scores for each pattern in the computer until the best item for inclusion in the subset was found, or to print out the pattern values of all subsets examined. On the other hand, the process of obtaining a mean score for each pattern was relatively easy and quick once all of the items in the subset were known.

The second part of the cross-validation program did in fact perform cross-validation. The program assigned the mean pattern scores from the validation sample to subjects having the same response patterns in the cross-validation sample and then correlated these scores with the final school grades of the 750 subjects in this sample.

The fact that all patterns may not have been assigned scores in the validation sample was handled by eliminating subjects from the cross-validation sample who had patterns that had not been assigned scores. This procedure was

considered acceptable because of the small number(eight) of subjects who fell into this catagory for a seven-item subset.

The cross-validation program provided the following information, given the items that form the subset:

1. A coded identification of the pattern of responses. APPENDIX B explains how to construct the patterns from the code.
2. The mean score for each pattern encountered in the validation sample.
3. An indication of which patterns of response were not encountered in the validation sample.
4. The validity of the validation sample.
5. The validity in cross-validation.
6. The number of subjects eliminated from the cross-validation sample because their patterns were not scored in validation.

The cross-validation program was also used to investigate what improvement in validity was obtainable through the use of PAIN over a random selection of a subset of items to use in pattern analysis.

The final form of the cross-validation program is contained in the COMPUTER PROGRAMS section(p. 31). For a seven-item subset this program had a running time of approximately ten seconds on the IBM 360 computer and used a core storage area of approximately 80K. APPENDIX A explains in detail how this program can be adapted for general use.

IV. RESULTS

PAIN selected items 20,14,40,56,7,5, and 33 in that order as the most predictive seven-item subset of the ETST. The validity of the seven items was .828 in validation for 1,500 subjects and .778 for the cross-validation of 750 subjects.

By using the pattern-analysis technique to score the subset of items selected by PAIN, it was possible to exceed the validity that had previously been attached to the ETST as a predictor of final school grades. The Navy had determined the predictive validity of the ETST to be approximately .61 using the total number of all 70 items correct as the predictor. A subset of as few as three items selected by PAIN was capable of establishing a predictive validity of .66 in cross-validation.

The cross-validation results for PAIN-selected items was an improvement over the .72 value of validity obtained in cross-validation by Lieutenant Weinberg in his study of SEQUIN's seven best items for predicting final school grades.

In the 15 cases investigated, the random selection of the first two items of the subset did not improve the validity of any six-item subsets (See Table 3). The items selected for the six-item subset under these conditions consistently included items selected by PAIN when no constraints were placed on the selection process. Eight

subsets selected only one item each that had not appeared in the unconstrained solution. Item number 16 was the only "new" item to appear in a six-item subset more than once, and it appeared in six different subsets.

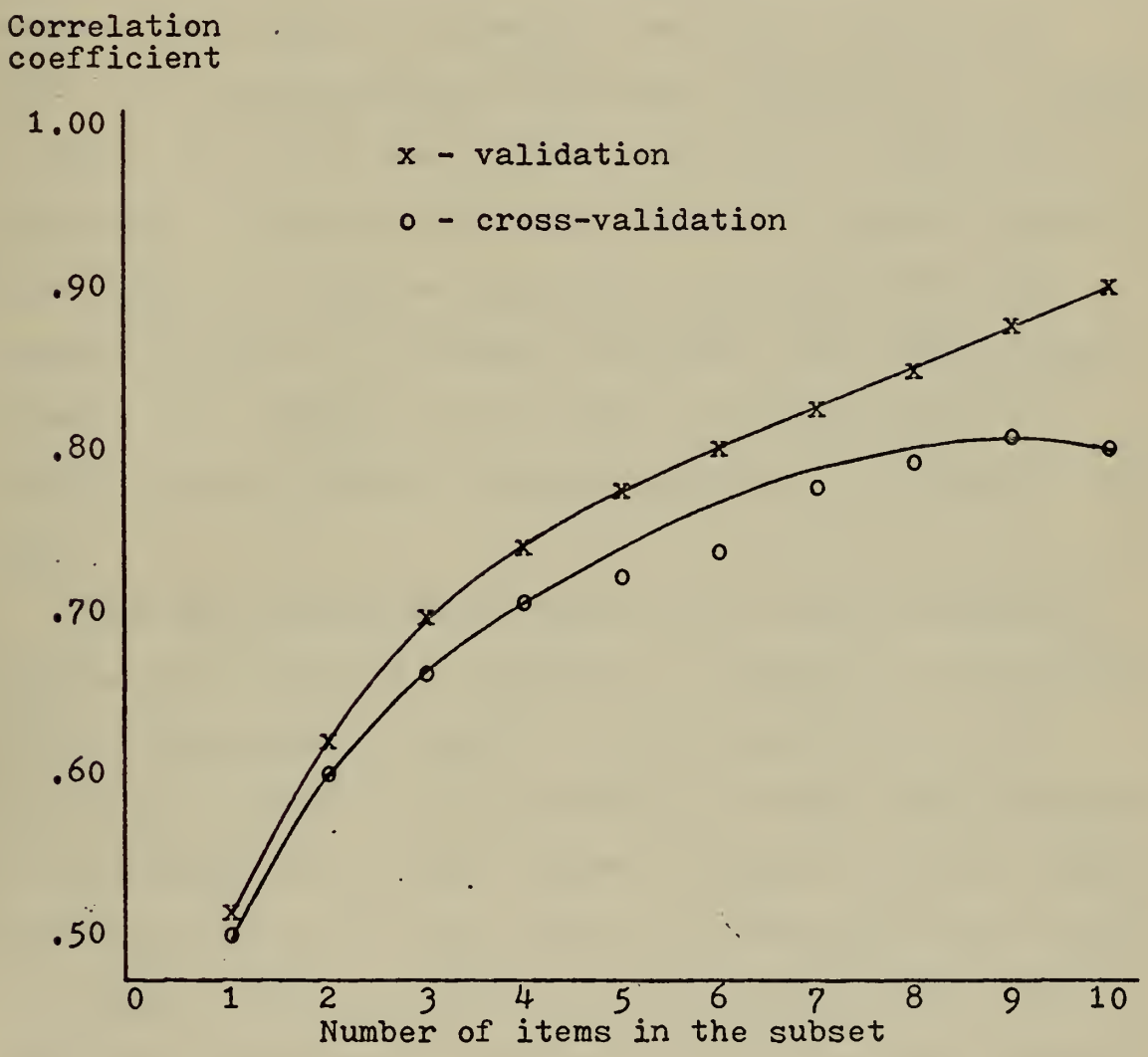
Random selection of seven items for pattern analysis also consistently resulted in lower validities than those obtained through the use of PAIN (See Table 4).

The author attempted using PAIN to select more than seven items from the ETST in order to determine at what point, if any, the validities in validation and/or cross-validation would level off or decline. At the ten-item-subset level, which was near the program size limit imposed by the computer system's core storage capacity, the validity was still increasing for the validation sample. The cross-validation sample, on the other hand, did show a decline in predictive validity at the ten-item level (See Figure 1, p. 17 and Table 1).

Examination of the assigned scores associated with the 128 patterns representing the best seven-item subset indicated that score assignments were not always directly related to the number of items correct in the subset. In some cases four correct items in the subset were assigned a lower score than three correct items. Also, the score assigned to getting only a certain item correct in a subset of one size was not always the same as the score assigned to getting only that item correct in a different size subset.

FIGURE 1

CORRELATION COEFFICIENTS IN VALIDATION AND CROSS-VALIDATION
OF
ITEMS SELECTED BY PAIN



Note: See Table 1 for exact values of validity coefficients

V. CONCLUSIONS

This study has in part confirmed the value of using pattern analysis as a predictive technique. While a method such as stepwise regression would have assigned a value to each of the items in the subset eventually selected by PAIN, the pattern-analysis approach allowed for the change in value of each of these items when used in different combinations. It appears that PAIN obtained a higher validity than SEQUIN or stepwise-regression techniques (MN 4112 student projects in progress concurrent with this writing) because more complete use was made of the information available when a pattern-analysis approach to evaluation was used.

PAIN and the scoring section of the cross-validation program have provided a predictor of Electronics Technician School final grades superior to any other known to the author at the time of this writing. By using this technique of item selection on other tests, a series of short, highly predictive tests for other areas requiring evaluation could be formed. A word of caution is appropriate though. The subjects of this study answered the questions used to form PAIN's seven-item subset while taking the 70-item ETST. The effect on the results of taking a seven-item versus 70-item test was not known at the time of this writing. Until it is determined that the shortness of the test is without adverse effect, the full implementation of testing based on subsets cannot proceed.

The consistent appearance of certain items in subsets formed with and without constraints on PAIN, coupled with the lower validities resulting when PAIN was constrained, would seem to indicate that the process of retaining items previously selected does not reduce the overall validity of a subset. Although the solution obtained by this method may not be a true optimal solution, it is questionable how much can be gained by an attempt at examining all possible solutions.

The author would have preferred to use a much larger sample than that used so that a much closer examination could have been made of the point at which subset size increases lack value. It is probable that the decline in validity in cross-validation at the ten-item level experienced in this study was a result of having less than two subjects available for each pattern of response. In fact, at the ten-item level over 13 percent of the cross-validation sample was unusable because of the lack of a scored pattern. The problem of availability and a desire on the author's part to avoid the possible problems associated with taking subjects who received final school grades based on differing grading systems caused the restiction on the size of the samples used.

A key area for more investigation is that of selecting predictors based on biographical information. Preliminary studies by others who have used PAIN (concurrent MN 4112

students) indicate that this technique of selection may have great value as a method of analyzing biographical data in relation to various criteria. This would seem logical if one will agree that patterns of information play a more important role in the area of biographical information than in the area of testing. APPENDIX C gives an explanation of how some biographical information can be converted into the binary form necessary for PAIN. APPENDIX C also contains examples of some of the preliminary results obtained by using PAIN in conjunction with biographical information.

TABLE 1

CORRELATION COEFFICIENTS FOR SUBSETS OF THE ETST
IN
VALIDATION AND CROSS-VALIDATION

NUMBER OF ITEMS IN SUBSET	VALIDATION	CROSS-VALIDATION
1	.51456	.50002
2	.62723	.60086
3	.69781	.66228
4	.74034	.70940
5	.77573	.72420
6	.80276	.73857
7	.82843	.77829
8	.85161	.78925
9	.87731	.81201
10	.90346	.80213

TABLE 2

PAIN PROGRAM RUNNING TIMES AND CORE STORAGE REQUIREMENTS
FOR
VARIOUS SIZE SUBSETS

NUMBER OF ITEMS IN SUBSET	APPROXIMATE RUNNING TIME	APPROXIMATE CORE STORAGE REQUIRED
7	6 Min.	270 K
9	9 Min.	310 K
10	13 Min.	360 K

Note: Figures presented are based on evaluating a 70-item test from a sample of 1,500 subjects. Running times are for IBM 360 computer.

TABLE 3

PAIN RESULTS WITH CONSTRAINTS PLACED
ON THE
FIRST TWO ITEMS OF A SUBSET

TWO ITEMS CONSTRAINED	ADDITIONAL ITEMS SELECTED	SUBSET VALIDITY
17, 3	40,14,56,7	.79603
6, 16	40,56,14,21*	.79058
19, 6	40,14,56,8*	.79074
1, 20	40,14,56,7	.79789
4, 2	20,40,56,7	.78515
39, 24	20,14,56,16*	.78450
37, 21	16*,14,40,56	.79395
36, 30	20,14,40,7	.77809
32, 29	14,40,7,56	.77285
34, 23	20,14,40,16*	.77978
55, 42	16*,14,40,7	.78773
54, 67	14,40,7,56	.76546
47, 63	14,40,7,56	.78603
56, 48	16*,14,40,7	.79528
69, 53	16*,40,14,20	.77085

Note: Additional items selected are presented in the order of selection by PAIN. The seven items originally selected by PAIN were 20,14,40,56,7,5, and 33 in that order.

*Indicates item not originally selected by PAIN

TABLE 4

VALIDATION AND CROSS-VALIDATION RESULTS
OF
PATTERN ANALYSIS TECHNIQUE
USED ON
RANDOMLY SELECTED SUBSETS OF SEVEN ITEMS

RANDOMLY SELECTED ITEMS	VALIDATION	CROSS-VALIDATION
12,14,38,41,42,48,56	.75235	.67070
4,40,58,63,66,67,70	.66504	.59745
15,23,40,48,54,58,59	.67265	.57414
4,10,17,26,34,45,55	.71356	.69899
6,14,41,42,53,56,66	.73571	.66944

Note: PAIN validation and cross-validation results for seven items were .82843 and .77829 respectively.

APPENDIX A

PAIN AND CROSS-VALIDATION PROGRAM CONVERSION FOR GENERAL USE

The PAIN and cross-validation programs in this study can be used to process any data of a binary nature simply by altering the contents of the DIMENSION and DATA statements and insuring that the READ statement conforms to the device from which the data is being read. This APPENDIX is a detailed check list of how the DATA and DIMENSION statements should be set up by the user.

A. PAIN DATA STATEMENT

1. N1 is set equal to the size of the sample being used in validation.

2. N2 is set equal to one (1). The program will handle increasing this variable to conform to the size of the subset under consideration.

3. N3 is set at a value equal to or greater than the integer range of the criterion scores. If the criterion scores are not in an integer form, conversion most be made before the data are read into the program so that the matrix involved can be addressed. A Data Conversion program can be altered to do this if such a program is used. Example: A 3.45 criterion score can be converted to a 345 criterion score. If conversion were not made in advance, the program would truncate this criterion score to 3. See

A7 in this Appendix for further details. The alternative to using integer criterion values would require a degree of manipulation within PAIN that is unwarranted for most cases.

4. N4 is set equal to two (2). The program will handle the increasing of this variable to conform to the number of patterns within a given subset size.

5. N5 is set equal to the final number of items desired in the subset.

6. N6 is set equal to the total number of items in the set under investigation.

7. INDEX is equal to a value one less than the lower number used in determining the range of the criterion (N3). Example: If the criterion were student grades on a 4.0 grading scale and the investigator did not know the actual value of the lowest grade in the sample, but knew the lowest grade was at least higher than 1.5, the value of N3 could be set as low as 40-15, or 25, and the value of INDEX would be 15-1, or 14. Note that if the lowest value for a final grade had actually been 1.5 instead of higher than 1.5 the conversion would have been 40-14, or 26, for the value of N3 and 14-1, or 13, for the value of INDEX. Although PAIN can be run without going through the process of assigning a value to INDEX, in many cases the core storage and running time of the program can be greatly reduced by using the INDEX variable.

B. CROSS-VALIDATION DATA STATEMENT

1. In order to perform cross-validation both the validation and cross-validation data must be read into the program respectively.

2. N1, N3, and INDEX follow the same rules as in the PAIN DATA statement.

3. N2 is set equal to the number of items in the subset being cross-validated or for which mean pattern scores are desired.

4. N4 is set equal to the number of patterns associated with the subset size being cross-validated. This value is equal to 2 to the N2 power.

5. N7 is set equal to the number of subjects in the cross-validation sample. Setting this value to zero (0) results in processing only the mean pattern scores for the validation sample.

C. DIMENSION STATEMENTS

1. The variables in the DIMENSION statements are dimensioned according to the comments at the beginning of PAIN and the cross-validation programs.

2. Definitions of the variables involved in the DIMENSION statements are contained in the list of non-dummy variables preceding the computer programs in the COMPUTER PROGRAMS section(p. 31).

APPENDIX B

INTERPRETING PATTERN CODES

All patterns of response were converted from binary to decimal form during PAIN and the cross-validation programs so that matrix addresses could be used. Hence, the list of patterns printed as output to the cross-validation program is in decimal form. The user of this program simply needs to subtract one from the pattern number in the output to obtain the actual decimal equivalent of the binary number of the pattern referenced. For example, the cross-validation program assigned a mean pattern score of 73.0 to the pattern listed as number 58. This would convert to the decimal equivalent 57, which yields the binary pattern 0111001. Since the items of the subset used were read into the cross-validation program in the order 5,7,14,20, 33,40,56, the pattern would indicate that a "correct" or "yes" answer to items 7,14,20, and 56 coupled with an "incorrect" or "no" answer to items number 5,33, and 40 predict a criterion score of 73.0.

APPENDIX C

BIOGRAPHICAL INFORMATION AND PAIN

To convert biographical information into the binary form necessary for use in PAIN requires questions to be formulated such that answers can be expressed in a yes-no form.

Questions that at first do not appear to fit a yes-no format are already being sectioned into parts so that that format can be used. For example, the question, "How old are you?", can be handled in the following way and often is:

How old are you?	Check one box
1. under 21	☐
2. 21 - 30	☑
3. 31 - 40	☐
4. over 40	☐

The boxes without checks can be considered as "no" answers in this example with the checked box a "yes". The one question, "How old are you?", can now be handled by PAIN as four separate items. If PAIN should indicate that one or more of these items represent good predictors, those items could be further sectioned for further evaluation. Other types of biographical information can also be handled in this manner.

The author knows of at least two studies, that were being conducted by students at the U. S. Naval Postgraduate

School at the time of this writing, which involved the use of PAIN for selecting items of a biographical nature for use in predicting various criteria. A study using the final QPA's of students in the Masters program at the U. S. Naval Postgraduate School as the criterion has yielded encouraging preliminary results. While samples were too small to justify comparison in cross-validation, PAIN provided uniformly higher validities than stepwise regression in validation.

The second study involved predicting drug addiction. This study had found a four-item subset that had a validity of over .60 in validation. No cross-validation results were available at the time of this writing, but any reasonable retention of validity in cross-validation could provide an extremely useful predictor in this field.

COMPUTER PROGRAMS

LIST OF NON-DUMMY VARIABLES USED IN COMPUTER PROGRAMS

ANS(I) - correct response to item I of test

B(J) - jth subject's assigned decimal value for his binary pattern of responxes

C(J) - jth subject's final school grade used for the criterion

C1 - sum of the criterion scores

C2 - sum of the squares of the criterion scores

D(J) - jth subject's identification number

F(M,N) - the joint frequency distribution of patterns versus criterion scores

INDEX - a value equal to the lowest score used in determining range (N3) minus 1

K - as used in the conversion program only, the number of the data card on which information is stored

M - row in the "F" matrix representing number of a pattern in a given subset

N - column in the "F" matrix representing the criterion score

N1 - the size of the sample used in validation

N2 - number of items in the subset being considered

N3 - coded range of the criterion scores

N4 - number of patterns in the subset being considered

N5 - size of subset desired

N6 - total number of items in set being investigated

N7 - the size of the sample used in cross-validation

P(I,J) - jth subject's binary response to item i

R2 - the correlation coefficient determined from the use of raw scores

S(I) - the mean pattern score for pattern i

S1 - sum of the criterion scores for a given pattern

S2 - sum of the subjects with a given pattern

W(I) - answer given by subject to item i of test

X(J) - jth subject's mean pattern score

X1 - sum of mean pattern scores

X2 - sum of squares of mean pattern scores

DATA CONVERSION PROGRAM

```
C THIS PROGRAM EDITS INFORMATION ABOUT SUBJECTS WHO HAVE
C TAKEN THE ETST, CONVERTS TEST DATA TO A BINARY FORM SUCH
C THAT A CORRECT ANSWER IS ASSIGNED THE VALUE '1' AND AN
C INCORRECT ANSWER IS ASSIGNED THE VALLE '0'. THIS
C INFORMATION IS THEN TRANSFERED TO A DATA CELL
C
      IMPLICIT INTEGER*4(A-Z)
      DIMENSION P(70),ANS(70),W(70)
      DATA NREAD,NWRITE,NPUNCH/8,9,7/
C
C THE CORRECT ANSWERS TO ALL ETST QUESTIONS ARE READ IN
C
      READ (5,1) (ANS(I),I=1,70)
    1 FORMAT (70I1)
C
C INFORMATION ON EACH SUBJECT IS READ IN
C
      DO 100 J=1,2400
      IF(J.EQ.2398) GO TO 50
   10 READ (NREAD,2,END=50) D,K,(W(I),I=1,70)
    2 FORMAT (T2,I6,I1,70I1)
      IF(K.NE.5) GO TO 10
      READ (NREAD,4) K,C
    4 FORMAT (T8,I1,T64,I2)
      IF(K.NE.6) GO TO 10
C
C CONVERT EACH SUBJECTS ETST ANSWERS TO BINARY FORM
C
      DO 20 I=1,70
      IF(W(I).NE.ANS(I)) P(I)=0
      IF(W(I).EQ.ANS(I)) P(I)=1
   20 CONTINUE
C
C OUTPUT THE EDITED AND CONVERTED INFORMATION
C
      IF(J.GT.44) GO TO 83
      GO TO 84
   83 IF(J.GT.150) GO TO 84
      WRITE (6,85) J,C,(P(I),I=1,70)
   85 FORMAT (I5,5X,I3,5X,7CI1)
   84 WRITE (NWRITE,8) D,C,(P(I),I=1,70)
    8 FORMAT (I6,I2,70I1)
  100 CONTINUE
   50 WRITE (6,9) D,C,(P(I),I=1,70)
    9 FORMAT (I8,5X,I2,5X,70I1)
      STOP
      END
```

PAIN: ITEM SELECTION PROGRAM

```
C THIS PROGRAM SELECTS A SUBSET OF ITEMS THAT MAXIMIZE
C VALIDITY UNDER PATTERN ANALYSIS. THE VALUES ASSIGNED TO
C  THE DIMENSIONED VARIABLES ARE: B(N1), C(N1), D(N1),
C F(2**N5,N3), P(N6,N1),S(2**N5),X(N1), ITEM(N5)
C
      INTEGER*2 B,C,F,P
      INTEGER C1,C2,R3,S1,S2
      DIMENSION B(1500),C(1500),D(1500),F(128,47),
     2P(70,1500),S(128),X(1500),ITEM(7)
      DATA N1,N2,N3,N4,N5,N6,INDEX/1500,1,47,2,7,70,29/
C
C THE SUBSET THAT WILL CONTAIN THE TEST ITEMS SELECTED IS
C INITIALIZED TO ZERO
C
      DO 11 I=1,N5
      ITEM(I)=0
   11 CONTINUE
C
C DATA IS READ INTO THE PROGRAM
C
      C1=0
      C2=0
      DO 13 J=1,N1
  350 READ (9,9) D(J),C(J),(P(I,J),I=1,N6)
    9 FORMAT (I6,I2,70I1)
C
C THE FOLLOWING TWO IF STATEMENTS PREVENT THE CONSIDERATION
C OF ANY SUBJECT WHO HAS A CRITERION SCORE OUTSIDE THE
C RANGE LIMITS USED IN ESTABLISHING N3 AND INDEX
C
      IF(C(J).LT.30) GO TO 350
      IF(C(J).GT.76) GO TO 350
      C1=C(J)+C1
      C2=C(J)*C(J)+C2
 13   CONTINUE
C
C THIS LOOP CONTROLS WHICH ITEM OF THE SUBSET IS BEING
C SELECTED DURING THE CURRENT ROUND OF EXAMINATIONS
C
      DO 220 L=1,N5
      RH=0.0
C
C THIS LOOP CONTROLS WHICH ITEM FROM THE TOTAL SET OF ITEMS
C IS BEING CONSIDERED FOR EXAMINATION
C
      DO 200 KA=1,N6
C
C THIS LOOP PREVENTS CONSIDERATION OF AN ITEM ALREADY
C SELECTED TO BE A MEMBER OF THE SUBSET
C
      DO 14 I=1,N2
      IF(KA.EQ.ITEM(I)) GO TO 200
   14 CONTINUE
C
C THE F MATRIX IS INITIALIZED TO ZERO
C
      DO 12 I=1,N4
      DO 17 J=1,N3
      F(I,J)=0
 17   CONTINUE
   12 CONTINUE
C
C THE JOINT FREQUENCY DISTRIBUTION OF PATTERN AND CRITERION
C SCORES IS DETERMINED AND BINARY PATTERNS ARE CONVERTED TO
C DECIMAL EQUIVALENTS TO BE USED AS ROW ADDRESSES.
C
```

```
      DO 18 J=1,N1
      M=1
      K=N4
      DO 19 I=1,N2
      K=K/2
      IF(ITEM(I).EQ.0) GO TO 19
      M=K*P(ITEM(I),J)+M
 19   CONTINUE
      M=K*P(KA,J)+M
      N=C(J)-INDEX
      F(M,N)=F(M,N)+1
      B(J)=M
 18   CONTINUE
C
C  THE MEAN CRITERION SCORES FOR EACH PATTERN ARE COMPUTED
C
      DO 20 I=1,N4
      S1=0
      S2=0
      DO 21 J=1,N3
      S2=F(I,J)+S2
      S1=(J+INDEX)*F(I,J)+S1
 21   CONTINUE
      IF(S2.EQ.0) GO TO 10
      S(I)=S1/S2
      GO TO 20
   10 S(I)=0.0
 20   CONTINUE
C
C  MEAN SCORES FOR PATTERNS ARE ASSIGNED TO EACH SUBJECT
C  ACCORDING TO HIS PATTERN
C
      DO 31 J=1,N1
      K=B(J)
      X(J)=S(K)
 31   CONTINUE
C
C  CORRELATION COEFFICIENT BETWEEN CRITERION AND PATTERN
C  FORMED USING ITEM UNDER CONSIDERATION IS COMPUTED.
C
      X1=0.0
      X2=0.0
      W=0.0
      DO 41 J=1,N1
      X1=X(J)+X1
      X2=X(J)*X(J)+X2
      W=C(J)*X(J)+W
 41   CONTINUE
      R1=(N1*X2)-(X1*X1)
      IF(R1.EQ.0.0) GO TO 85
      R3=(N1*C2)-(C1*C1)
      R5=(N1*W)-(C1*X1)
      Q=(R1*R3)**0.5
      R2=R5/Q
      GO TO 86
   85 R2=0.0
C
C  IT IS DETERMINED IF THE CORRELATION COEFFICIENT USING
C  ITEM PRESENTLY UNDER CONSIDERATION IS HIGHER THAN HIGHEST
C  PREVIOUSLY FOUND CORRELATION COEFFICIENT WITH SAME SIZE
C  SUBSET. ITEM NUMBER AND CORRELATION COEFFICIENT ARE
C  STORED IF THEY ARE HIGHER.
C
   86 WRITE (6,54) KA,R2
   54 FORMAT (' THE CORRELATION OF ITEM ',I4,' WITH THE
     2PREVIOUSLY SELECTED ITEMS IS',F10.9)
      IF(R2.GT.RH) GO TO 81
      GO TO 200
   81 RH=R2
      ITEMH=KA
  200 CONTINUE
C
```

```
C PRINT CUT THE ITEM NUMBERS OF ITEMS USED TO GET THE
C HIGHEST CORRELATION COEFFICIENT FOR THE SIZE SUBSET UNDER
C CONSIDERATION AND THE VALUE OF THIS COEFFICIENT.
C
      ITEM(L)=ITEMH
      WRITE (6,160) L,L,(ITEM(I),I=1,N5),RH
  160 FORMAT ('0','THE BEST ',I2,'ITEMS TO USE FOR A ',I2,
     2' ITEM SUBSET ARE:',/' ',10I5,' AND THEY YIELD A
     3CORRELATION COEFFICIENT OF ',F10.9,//)
C
C ADVANCE THE NUMBER OF ITEMS IN THE SUBSET AND THE NUMBER
C OF PATTERNS POSSIBLE THEN REPEAT THE EXAMINATION PROCESS.
      N2=N2+1
      N4=2**N2
  220 CONTINUE
      STOP
      END
```

CROSS-VALIDATION PROGRAM

```
// EXEC FORTCLG
C THIS PROGRAM PRINTS THE MEAN PATTERN SCORES FOR THE
C SUBJECTS IN THE VALIDATION SAMPLE AND CROSS-VALIDATES THE
C CROSS-VALIDATION SAMPLE. THE VALUES ASSIGNED TO THE
C DIMENSIONED VARIABLES ARE:B(N1), C(N1),F(N4,N3), P(N2,N1),
C S(N4), X(N1)
      INTEGER*2 B,C,D,F,P
      INTEGER*4 C1,C2,R3,S1,S2
      DATA N1,N2,N3,N4,N7,INDEX/1500,7,47,128,750,29/
C
C DATA IS READ IN FROM THE INPUT DEVICE
C
      DO 220 L=1,2
      IF(L.EQ.2) N1=N7
      IF(N1.EQ.0) GO TO 900
      DO 13 J=1,N1
  350 READ (9,9) C(J),(P(I,J),I=1,N2)
    9 FORMAT (T7,I2,T13,I1,T15,I1,T22,I1,T28,I1,T30,I1,
     2T48,I1,T64,I1)
C
C THE FOLLOWING TWO IF STATEMENTS PREVENT THE CONSIDERATION
C OF ANY SUBJECT WHO HAS A CRITERION SCORE OUTSIDE THE
C RANGE LIMITS USED IN ESTABLISHING N3 AND INDEX
C
      IF(C(J).LT.30) GO TO 350
      IF(C(J).GT.76) GO TO 350
   13 CONTINUE
      RF=0.0
C
C THE F MATRIX IS INITIALIZED TO ZERO
C
      DO 12 I=1,N4
      DO 17 J=1,N3
      F(I,J)=0
   17 CONTINUE
   12 CONTINUE
C
C THE JOINT FREQUENCY DISTRIBUTION OF PATTERNS AND
C CRITERION SCORES IS DETERMINED AND BINARY PATTERNS ARE
C CONVERTED TO DECIMAL EQUIVALENTS TO BE USED AS ROW
C ADDRESSES
C
      DO 18 J=1,N1
   29 M=1
      K=N4
      DO 19 I=1,N2
      K=K/2
      M=K*P(I,J)+M
   19 CONTINUE
      N=C(J)-INDEX
      F(M,N)=F(M,N)+1
      B(J)=M
   18 CONTINUE
      IF(L.EQ.2) GO TO 200
C
C THE MEAN CRITERION SCORES FOR EACH PATTERN ARE COMPUTED
C AND PRINTED OUT.
C
      WRITE (6,94)
   94 FORMAT (' ',5X8'PATTERN NUMBER',10X,'MEAN CRITERION
     2SCORE',//)
      DO 20 I=1,N4
      S1=0
      S2=0
      DO 21 J=1,N3
      S2=F(I,J)+S2
```

```
      S1=(J+INDEX)*F(I,J)+S1
 21   CONTINUE
      IF(S2.LT.1.0) GO TO 10
      S(I)=S1/S2
      WRITE (6,93) I,S(I)
   93 FORMAT (' ',11X,I3,20X,F12.5)
      GO TO 20
   10 S(I)=0.0
      WRITE (6,90) I
   90 FORMAT (' NO SUBJECT HAD THE FOLLOWING PATTERN DURING
     2'ITEM SELECTION PROGRAM:',I4)
 20   CONTINUE
  200 CONTINUE
C
C MEAN SCORES FOR PATTERNS ARE ASSIGNED TO EACH SUBJECT.
C
      NR=0
      DO 31 J=1,N1
      K=B(J)
      X(J)=S(K)
      IF(S(K).LT.0.9) GO TO 91
      GO TO 31
   91 NR=NR+1
      C(J)=0
 31   CONTINUE
      IF(L.EQ.1) GO TO 32
      WRITE (6,92) NR
   92 FORMAT (I6,' SUBJECTS HAVE PATTERN SCORES NOT
     2ENCOUNTERED DURING THE MAIN PROGRAM')
C
C THE CORRELATION COEFFICIENT IN VALIDATION AND
C CROSS-VALIDATION FOR THE ITEMS UNDER CONSIDERATION IS
C COMPUTED AND PRINTED OUT.
C
   32 C1=0
      C2=0
      X1=0.0
      X2=0.0
      W=0.0
      DO 41 J=1,N1
      C1=C(J)+C1
      C2=C(J)*C(J)+C2
      X1=X(J)+X1
      X2=X(J)*X(J)+X2
      W=C(J)*X(J)+W
 41   CONTINUE
      N1=N1-NR
      R1=(N1*X2)-(X1*X1)
      R3=(N1*C2)-(C1*C1)
      R5=(N1*W)-(C1*X1)
      Q=(R1*R3)**0.5
      R2=R5/Q
   81  RH=R2
      IF(L.EQ.2) GO TO 220
  160 FORMAT ('0','THE ITEMS UNDER CONSIDERATION YIELD A
     2VALIDITY OF ',F10.5,' IN VALIDATION')
  220 CONTINUE
      WRITE (6,54) N2,RH
   54 FORMAT (' THE VALIDITY OF THE',I2,'ITEMS CONSIDERED
     2FOR CROSS-VALIDATION IS ',F10.5
  900 STOP
      END
```

BIBLIOGRAPHY

1. U. S. Naval Personnel Research Activity Research Memorandum SRM 67-8, SEQUIN: A Computerized Item Selection Procedure, by W. J. Moonan and CPL. U. W. Pooch, USMC, October 1966

2. Naval Personnel and Training Research Laboratory Technical Bulletin STB 70-3, A Preliminary Evaluation of Brief Navy Enlisted Classification Tests, by L. Swanson and B. Rimland, January 1970

INITIAL DISTRIBUTION LIST

	No. Copies
1. Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2. Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3. Professor R. A. Weitzman Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	2
4. LCDR. Burton F. Folce, Jr., USCG (student) U. S. Coast Guard Headquarters (PE) Washington, D.C. 20590	1
5. Commandant (PTP) U. S. Coast Guard 400 Seventh Street SW. Washington, D.C. 20590	2
6. Director of Admissions U. S. Coast Guard Academy New London, Connecticut 06320	1
7. Test Evaluation Division U. S. Coast Guard Institute Oklahoma City, Oklahoma 73102	1
8. Naval Personnel and Training Research Laboratory San Diego, California 92152	1
9. Library (Code 55) Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Naval Postgraduate School Monterey, California 93940	Unclassified
	2b. GROUP

REPORT TITLE

Development of a Pattern Analysis Technique for use in the Selection of Predictors

DESCRIPTIVE NOTES (Type of report and inclusive dates)

Master's Thesis; June 1973

AUTHOR(S) (First name, middle initial, last name)

Burton F. Folce, Jr.

REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1973	42	2

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

This study describes a computerized item-selection program called PAIN that uses a pattern-analysis approach to select a most-valid subset of items from a set. The results of this study indicate that PAIN is capable of selecting a small subset of items which, when scored by pattern analysis, has greater validity than the original set. It appears that, as well as reducing the sizes of standard tests without losing predictive value, PAIN may also be of value in selecting biographical items of information for use as predictors.

DD FORM 1 NOV 68 1473 (PAGE 1)

S/N 0101-807-6811

UNCLASSIFIED
Security Classification

A-31408

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Pattern Analysis						
SEQUIN						
PAIN						

Thesis 144177
F535 Folce
c.1 Development of a pattern analysis technique for use in the selection of predictors.